On-line Context Aware Physical Activity Recognition from the Accelerometer and Audio Sensors of Smartphones

David Blachon, Doruk Cokun, François Portet

On-line Context Aware Physical Activity Recognition from the Accelerometer and Audio Sensors of Smartphones

David Blachon[1,2], Doruk Coşkun[1], and François Portet[1]

[1] Laboratoire d'Informatique de Grenoble
Univ. Grenoble Alpes, LIG, F-38000 Grenoble, France
CNRS, LIG, F-38000 Grenoble, France
41 rue Mathématiques, BP 53, 38041 Grenoble cedex9, France
{david.blachon,doruk.coskun,francois.portet}@imag.fr
[2] STMicroelectronics, 12 rue Jules Horowitz, BP 217, 38019 Cedex, Grenoble, France
david.blachon@st.com

Abstract. Activity Recognition (AR) from smartphone sensors has become a hot topic in the mobile computing domain since it can provide services directly to the user (health monitoring, fitness, context-awareness) as well as for third party applications and social network (performance sharing, profiling). Most of the research effort has been focused on direct recognition from accelerometer sensors and few studies have integrated the audio channel in their model despite the fact that it is a sensor that is always available on all kinds of smartphones. In this study, we show that audio features bring an important performance improvement over an accelerometer based approach. Moreover, the study demonstrates the interest of considering the smartphone location for on-line context-aware AR and the prediction power of audio features for this task. Finally, another contribution of the study is the collected corpus that is made available to the community for AR recognition from audio and accelerometer sensors.

Keywords: Data Science, Sensing and Reasoning Technology

1 Introduction

Automatic human Activity Recognition (AR) is recognised as an important process for human behaviour monitoring (health, well-being, sport) [8] but it is also extensively studied for the provision of context-aware services for smart objects (smartphones, robots...) and smart spaces (smart homes, smart rooms, public spaces...) [7]. Though AR is a hot topic in the community, according to the application, the activities under study can be very different (e.g., walking, getting money at an ATM, screw driving, etc.) and there is still not a clear definition of the different levels of activities that can be modelled and processed. In this article, we aim at detecting basic physical activities from smartphone i.e., activities involving basic movements with low level of semantics such as walking, standing

still or sitting. These are a subset of the Compendium of Physical Activities [1] that are classified according to the semantics they share and effort magnitude. For instance, the group *Bicycling* lists different activities of bicycling in several environments (e.g., mountain) and at different speeds. Basic physical activities have the advantages of being well defined, can be captured by means of sensors and can be used to reconstruct higher-level activities. They are also directly related to the health and sporting activities of the user and can greatly inform the context of smartphones.

The emergence of smartphones introduced new ways to perform human activity recognition. Indeed, they embed accelerometers for sensing, and resources for storage and data transmission. Hence, data collection is comparable to previous studies such as the one of Bao et al [3]. Beyond the availability of accelerometer, smartphones usually embed many different sensors. For instance, it is common to find other inertial sensors (e.g., gyroscope), ambient sensors (e.g., microphone, barometer, magnetometer) which can be used for human activity recognition. Also, thanks to their design, smartphones are easy to carry so they might be used to collect new kinds of activities in realistic conditions. Studies report their use for tasks about locomotion activities [12, 17] or daily activities [10].

Smartphones also embed large resources in terms of computation, storage, battery, which could allow to perform online embedded activity recognition. However, according to a survey from Incel et al [11], online activity recognition is an under explored area. They report that most studies deal with offline classification and that classifiers still require much resource for embedding them on smartphones. Yet, some recent work have started studying online AR classification, using Decision Tree and K-Nearest Neighbors (KNN) [17, 12].

Beyond those temporary limitations of resources, other issues need to be tackled with. First, the large variability of sensors does not seem to be standardized yet, which means that one should not make a system depend on the availability of such a sensor on every smartphone. For instance, accelerometer is quite common but proximity sensor or barometer are far less common. Hence, a system of human activity recognition should deal with this variable sensor availability. Also, unlike the previously mentioned study of Bao et al [3], sensor location and orientation may change in time. Indeed, smartphone users can carry them in different locations. A survey [6] performed among 55 volunteers reported the preferred locations for users to carry their smartphones. The 4 most frequent answers were hand, pants pocket, bag and jacket pocket. The change of location may make it more difficult for a system to infer user activity, yet it needs to be taken into account.

In this paper, we present an approach for online basic physical AR from microphone and accelerometer data streams collected on smartphones. Indeed, these two kinds of sensors are found on most smartphones (if not all). If accelerometers are very popular, microphones are far less used in the domain despite their strong informative potential. Another contribution of the study is the explicit modelling of the smartphone location context and its use as input information for AR. This AR framework is described in detail in Section 3 after a

short description of the state of the art in Section 2. Since no dataset composed of microphone and accelerometer data is available, we collected our own dataset on different smartphone brands to evaluate the method. This data collection, that we made publicly available, is described in Section 4. Three different state of the art classification models were learned from this dataset in different conditions to assess the impact of the audio channel and the smartphone context on classification performance. These results are detailed in Section 5 and discussed in Section 6. The paper finishes with a short conclusion and an outlook on further works in Section 7.

2 Related work

A first study dealing with basic physical activities is the one of Kose et al [12]. Authors report a high F-Measure value of 92% for a clustered KNN classifier on a set of basic physical activities (running, walking, standing, sitting). Another study is from Yang et al [21] focusing on a similar set of basic physical activites and using different classifiers such as Decision Tree, Naive Bayes, KNN and Support Vector Machine (SVM). This study reports 90% of accuracy using the Decision Tree. Moreover, the survey of Incel et al [11] shows that most reported studies rely on the use of accelerometer. GPS and radio communications such as WiFi are sometimes used while audio is hardly present in studies. Also, the survey shows that for most studies reported, the number of subjects was less or equal than 20.

Accelerometer appears to be the most popular sensor for the domain. However, we previously noticed in the introduction that smartphone location and orientation might change due to the habits of users that can carry it in different locations. This can have an impact on accelerometer readings as Alanezi et al [2] report. They collected accelerometer data from two different positions: hand holding and pants pocket. Magnitudes of acceleration hardly reached the value of 15 m/s^2 when in hand while they often exceeded 15 m/s^2 and even reached 20 m/s^2 in pants pocket. The difference was also noticeable on standard deviations of readings. Hence, as the authors concluded, accelerometer data are affected by smartphone position. However, studies reported different solutions to address this issue. First one is to train classifiers for each different location considered. Reddy et al [17] trained Decision Tree cascaded with Hidden Markov Models (HMM) for each different location and compared results to the same type of classifier trained with data from all concerned locations. They report similar performances for transportation mode recognition. Another reported technique includes a gravity-based feature for estimating orientation. Park et al [15] used this technique for recognizing different smartphone locations. Using a SVM classifier, they reported accuracy values of 99.6% when including the gravity-based feature and only 82% when excluding it. Despite those solutions, Incel et al consider in their review [11] that smartphone location and orientation still remain open points.

Leveraging the many sensors of smartphone can also help for activity recognition. Among them, one rarely used is the microphone that could be a source of information very relevant regarding such conditions. Indeed, a microphone can be found on every smarpthone. Although audio readings might suffer from frictions between microphone and direct context (e.g., fabric of pants) as reported Alanezi et al [2], some studies have promising results with microphone. For instance, Miluzzo et al [14] proposed a system to distinguish various smartphone locations thanks to environment sounds but it was successfully tested only on two locations: inside a pocket and outside a pocket. Diaconita et al [9] presented a technique to recognize smartphone position and user's environment based on the emission of a sound pulse and then the classification of the reverberation (classifiers were trained on such reverberations in order to recognize user's environment and smartphone position). They report 97% of accuracy using Mel Frequency Cepstral Coefficients (MFCC) as features classified with KNN. However, to the best of our knowledge, the use of audio for smartphone location recognition and human activity recognition is not optimal and can still be improved.

3 Method

This section describes the method designed for online activity recognition. The activities of interest as well as the smartphone context are first defined, then the global architecture, the extracted features and the different classification schemes are introduced.

3.1 Activities and Smartphone Context

In this article we focus on basic physical activities such as walking, running, etc. leaving apart high level activities (e.g., making coffee, cleaning the house) and fine grained gestures (e.g., screw driving, raising the phone over the head).

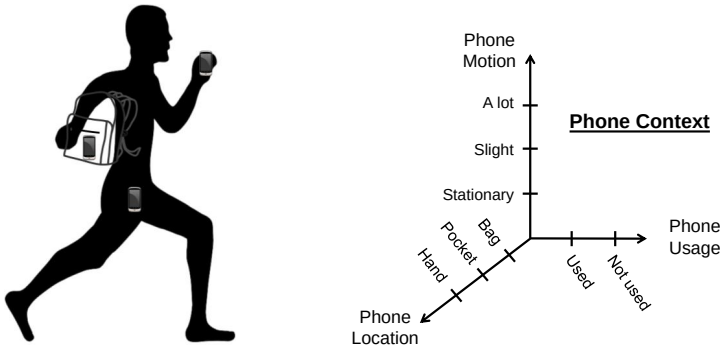
The activity set is presented Table 1. It is composed of i) 8 frequent daily activities which are a subset of the Compendium of Physical Activities [1] and ii) an Unstable activity which groups every data not belonging to any of the 8 previously defined activities (e.g., making coffee, cleaning). In any of these daily activities, the user is likely to wear his/her smartphone. This cannot be assumed in some other activities of the compendium (e.g., snorkeling).

Table 1. The 8 activities considered in the study and the UNSTABLE class.

Stationary	Sitting	Standing Still
Lying	Running	Walking
Jumping	Stairs	UNSTABLE

Though the above activities are well defined, the observations recorded by a smartphone can be highly dependent on the location of this device with respect

to the user. For instance, as exemplified by Figure 1(a), if someone is running, the observations collected will be very different if the smartphone is in the hand (in that case most of the dynamic will be those of the arm), in the bag (in that case the smartphone captures the bag jerks) or in the pocket (in that case most of the dynamic will be those of the legs). To take the different interaction dimensions between the user and the smartphone into account, we defined the smartphone context space represented Figure 1(b).



(a) Smartphone locations considered (b) Phone conext modelling

Fig. 1. Smartphone locations considered and smartphone context modelling

The first axis concerns phone motion. It is indeed very informative to assess which amount of movement the smartphone is subject to. This can help infer activity as well as whether the smartphone is with the user or not (e.g., put on a table far away from the user). The second axis, phone usage, has two values respectively used and not used. By phone usage we focus on the interaction with the phone. For example, typing an SMS is a direct interaction and listening to music is an indirect one, but both are considered as 'used' in our definition. On the opposite, carrying the phone in a bag without interacting with it is considered as not used. This information is a useful context component to build applications that can prompt user when she/he is paying attention to her/his smartphone. The third axis, smartphone location, represents the different locations in which a smartphone can be with respect to the user's body. For the kind of activities addressed in the study this is the most important information, therefore throughout the rest of the paper we will concentrate on the dependence between AR and smartphone location.

3.2 Architecture

The AR system architecture is depicted Figure 2. Data streams are continuously acquired. Every w windows, a buffer is sent for feature extraction. Features are computed over a 2-second long buffer with 50 % overlapping. From the feature

vector obtained, the most likely location of the smartphone is extracted. Then, the location and the feature vector feed the AR module which decides which activity (if any) is observed from the data.

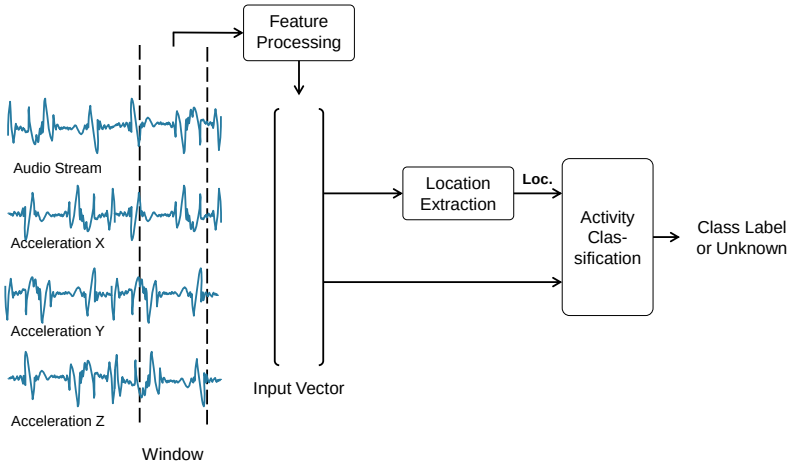


Fig. 2. Architecture of the online activity recognition system on smartphone

3.3 Features

Our set includes features from acceleration and audio data. They are listed below:

- Average of accelerometer magnitude
- Variance of accelerometer magnitude
- Energy of accelerometer magnitude
- Variance of acceleration along each axis (x, y, z)
- Subband energy at 3 Hz and 4 Hz

Average, variance and energy are very popular features for activity recognition task [11]. Since acceleration is directly related to motion, they estimate quantity and variability of motion within the window. Also, frequency parameters can help focus on human motion and avoid some noise added by data collection.

Regarding audio features, we used frequency based one but no temporal-based ones. Indeed, microphone is an ambient sensor that records data coming from different sources, including many external ones (e.g. radio, speech). Separating information due to motion from the one due to external source is a challenging task, if not impossible, using temporal domain features. However, frequency domain features can leverage energy distribution along frequencies allowing then to isolate spectrum subbands more sensitive to motion variation. For instance, as we reported in Section 2, a smartphone located in a pocket may involve frictions on the microphone which can in turn generate sounds. Finding

subbands of spectrum of these frictions could be very helpful. Our set of audio features contains spectrum power estimation of 40 subbands spread along the Mel scale frequencies. Mel-scales are inspired from human ability of hearing and they are basically perceptual scales of pitches judged by listeners to be equal in distance from one another. In theory, the computation of Mel filter banks is implemented in 3 steps 1) the signal is windowed (e.g., with a hamming window), 2) Fast Fourier transform is applied, 3) spectrum power estimation obtained is mapped onto the mel scale. For the experiments presented in this paper, spectrum power estimation is performed on 200 ms long frames and only spectrum magnitude coefficients are saved (phase is removed). From this audio feature set, speech content cannot be retrieved with today techniques.

3.4 Classification Models

Three different models were implemented to perform AR. Decision trees are the standard method used in the literature and was used as baseline system. Random Forest is the scheme that has gained interest recently since it showed the most promising performance in AR from smartphone and worn sensors [8]. The third model is Conditional Random Field (CRF) which is a sequence model which has demonstrated impressive performance in related field of AR [20]. These three models and their learning strategy are briefly introduced below.

Decision Trees The induction of a decision tree is based on the "divide and conquer" principle to partition a training set TS, composed of individuals described by several attributes, into homogeneous subgroups. Let the classes of the individuals be $\{C_1, C_2, \ldots, C_k\}$, there are four possibilities:

1. TS contains individuals belonging to only one class C_j. In this case, this is a leaf named C_j.
2. TS is empty. In this case this is a leaf for which the class is defined by information other than TS (e.g. the most frequent class of the parent).
3. TS contains individuals belonging to several classes. Thus, a test T is chosen, based on a single attribute that has the exclusive outcomes $\{O_1, O_2, \ldots, O_n\}$ which are used to partition TS into the subsets $\{TS_1, TS_2, \ldots, TS_n\}$ where TS_i contains all the individuals in TS that have outcomes O_i.The decision tree for TS consists of one branch for each outcome. This mechanism is then reapplied recursively on each subset TS_i .
4. TS contains individuals belonging to several classes but for which no test can be found. In this case this is a leaf for which the class is defined by information from TS (e.g. the most frequent class in TS) or other than TS.

In this paper, we use the well known C4.5 method [16] which uses the gain ratio to choose the test T. The gain ratio can be described as the gain of information (based on the entropy) for T normalized by the potential information of dividing TS into n outcomes. Therefore, the decision tree chooses the most discriminant tests (best separation of the information).

Random Forest Random forests is an ensemble classifier composed of several 'weak' decision trees whose each individual output are combined (by voting strategy) to generate the final decision. The method to induce Random Forest models [5] combines bagging and random attribute selection. Briefly, if the dataset contains N individuals described by M attributes, the learning consists in drawing a subset of individuals randomly with replacement to constitute a training set (keeping the rest as test set) which is used to induce a decision tree with m attributes ($m << M$) randomly selected. Each tree is grown to the largest possible extent without pruning. The most important parameters for the learning are the maximum number of trees and m the number of attributes. Random forest is very efficient on large datasets and has a very good accuracy on many datasets without over-fitting (even if some study downplayed this advantages with some noisy datasets [18]). In fact, RF has shown impressive performances for AR on worn sensors [8].

Conditional Random Fields Conditional random fields (CRF) are graph based models to perform discriminative probabilistic inference over a set of variables in classification tasks [13]. Similarly to HMM, CRF can classify a label y for a given sample x taking into account its relations with neighbour samples when they are organized in a sequence $X = x_1, \ldots, x_T$. CRF model the conditional distribution $p(Y|X)$, but without requiring to model the distribution of the variable X. In the case of activity recognition, we can consider the X a set of temporal windows, and Y as the activity to infer.

Formally, a CRF model is a undirected graph $G = (V, E)$ such that $Y = (Y_v)_{v \in V}$, so that Y is indexed by the vertices of G. Then (X, Y) is a conditional random field if the following Markov property with respect to the graph is respected $p(Y_v|X, Y_w, w \neq v) = p(Y_v|X, Y_w, w \sim v)$, where $w \sim v$ means that w and v are neighbours in G. The simplest of the CRF model is a simple chain. This is the model that we adopt in this paper. In that case, X represents a sequence of observations and Y represents a hidden state variable (activity) that must be inferred given X.

CRF are generally implemented as log linear models by means of feature functions f_k, in the case of chain-like conditional random fields $p(Y|X)$ is estimated by equation 1.

$$p(y|x) = \frac{1}{Z(x)} exp \left\{ \sum_{k=1}^{K} \lambda_k f_k(y_t, y_{t-1}, x_t) \right\} \quad (1)$$

where the features functions f_k impose a bias in the model, they take the value of 1 if its variables take a certain configuration and 0 otherwise, y_t is the class (activity) to be estimated at current time, x_t is the observation sequence at current time, y_{t-1} is the (previous) neighbour, Z is a normalization factor, λ_k is a parameter to assign a weight to the feature function f_k. These weights are estimated during the learning phase using an iterative gradient method. Finally, for chain-like model, inference is performed through an adapted version of the Viterbi algorithm [19].

4 Experiment

To validate the approach, datasets containing audio and accelerometer data labelled with physical activities were required. However, although a fair number of datasets with accelerometer data are available, we did not find a dataset including audio data. Therefore we ran our own experiment to collect data from 19 people by using 4 smartphones which were located in 3 different locations. This section describes the method employed to collect this dataset, the specific scenarios that were designed and finishes with a summary of the collected data.

4.1 Protocol

Each recording involved one experimenter and one participant as well as five smartphones. At the beginning of the experiment, the participant was fit with 2 smartphones in a bag, 1 in the pocket and 1 in the hand. Each participant performed the 9 activities introduced in Table 1 (walking, running, jumping, stairs, lying, standing still, stationary, sitting and the remaining one we call unstable) and 5 different secondary activities (calling, listening to music, sending sms, using an application on the phone, none) with predefined order which was determined by a data collection scenario. During that time, the experimenter was always with the participant to trigger changes of activities with his own smartphone. The own smartphone of the participant was never used. All smartphones were equipped with the RecordMe application[3] that acquires all sensors and smartphone activity data on the fly. Accelerometer data were sampled at a rate of 50 Hz and audio sampled at a rate of 44.1 kHz. More details of the RecordMe application can be found [4].

If the participant did not have a bag (e.g., own handbag) a backpack already filled with some stuff was used (e.g., pens, thermos bottle, sheets of paper). All the experiments happened within the lab in different places (office, corridors, stairs, close outside, cafeteria) at calm and rush hours. Before experiment, the participant was explained the aim of the study, the risk of undertaking this experiment so that she/he could give a signed informed consent to undertake the experiment and to let the data be used for research purpose. It was also explained that the raw audio data were not kept but only coefficients from which raw signal cannot be reconstructed.

4.2 Scenarios

To make sure all activities and smartphones will be uniformly mixed, our data collection consisted of 3 main scenarios. The organisation of each scenario was implemented by using the modelling of the smartphone context and human activities. All possible smartphone contexts and human activities that can be performed within relevant protocol were indicated. Table 2 summarises these 3 scenarios.

[3] `http://lig-membres.imag.fr/blachon/download/recordme.apk`

Table 2. Abstracted view of the scenarios used for data collection

scenario ID	Phone Usage	Phone Motion	Smartphone Location				Activities
			SP1	SP2	SP3	SP4	
1	Yes/No	Yes/No	bag	hand	pocket	bag	All activities + secondary act.
2	No	No	pocket	bag	hand	bag	Stationary
3	No	Yes/No	hand	pocket	bag	bag	All activities

In scenario 1, the participant is interacting with the phone (also carrying it). She/he has a phone in the hand (another one in the pocket and two in the bag) and performs all activities one after the other. To add naturalness to the collected data, the scenario included 'secondary activities'. These are performed with the smartphone in hand and contain activities such as calling or sending a SMS. These were added for representing the changeable hand movement of the participants. In daily life, people can carry their phones in their hand with different gestures. These gestures were modelled by the secondary activities. Then, in scenario 2 the participant is not carrying the phone in the hand. Instead, s/he puts it on a table and stays around quietly. While in scenario 3 the participant is carrying the phone (but not interacting with it) and perform again physical activities, most of them outdoors.

These 3 scenarios were performed sequentially without any interruption. Unlike many data collections in which activities are collected separately from each other, we wanted to reproduce the real life in which activities are part of a whole day. Indeed, systems to be deployed in the wild must deal with transitions between activities and not predefined activities.

A 'phone switch' sequence was inserted between the execution of each scenario so that smartphones were not always at the same place. Initial position of smartphone was uniformly distributed over the participants to secure the same amount of data per location $\times$ activity. That is to say, smartphones SP1-3 were not recording an activity always in the same location.

4.3 Acquired Data

At the end of the experiment, 20 hours of data were collected from 19 participants (14 males, 5 females) between 19-29 years old. Four smartphones were used: Motorola Defy Mini, MTT, Samsung Galaxy S2 and Nexus 4. Each experiment took 15 to 20 minutes to be performed. Figure 3 displays time distribution over the activities. Repartition of recorded time per smartphone is of 50 % for bag location, and 25 % for pocket and hand locations. All the data were annotated according to two dimesnions: Phone Location, and physical Human Activity. This dataset is available at `anonymous`

After data collection, the annotation performed during the experiment was manually checked in the time and label domains. Some data are still not cleaned and checked.

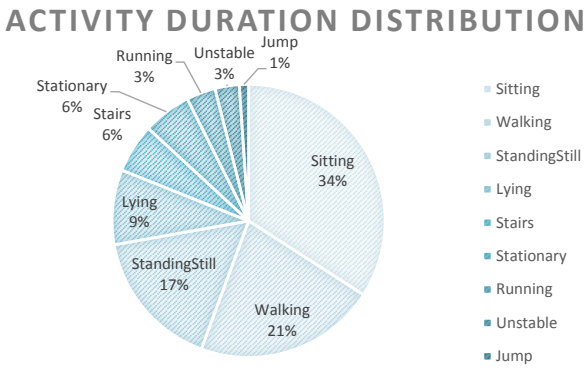


Fig. 3. Activity duration distribution over all 9 activities

5 Results

Model training and tests were performed offline. However, the classification task simulated online conditions since only the present data vector was known by the classifier. The tasks of activity recognition and smartphone location recognition were run on 3 different feature sets: 1) Accelerometer features; 2) Audio features; and 3) Accelerometer plus audio features. We used 50 trees for Random Forest. All models were acquired and tested using a 10 times x 10 fold cross validation and precision, recall and F-Measure were computed for each run. For each result, statistical significance test were computed against either the Decision tree (reference model) or accelerometer results (reference feature set).

5.1 Basic Human Activity Recognition

The first experiment was to test how well the models perform on the three different feature sets without using smartphone location information. Table 3 summarises the F-measure obtained for each combination. It turned out that Random Forest performed better than the two other algorithms in every feature sets. This increase of performance was for each set significantly better than Decision Tree. On the contrary, CRF performed worse than Decision Tree. Also, one can notice that Random Forest performed better with audio feature set than with accelerometer set. A significance test was performed to test whether audio features play a role in the performance increase. Each Model performance for the Accel+Audio feature set was compared to the Accel feature set. Apart for CRF model, improvements were significant ($p < 0.01$). Thus, it seems that audio features may play an important role for physical activity classification from smartphone.

5.2 Context-aware Human Activity Recognition

To assess the impact of smartphone location on activity classification, the learning has been run again but adding the ground truth location in the feature

Table 3. Overall F-measure of activity classification for each Model and feature set without smartphone location

Feature set	Decision Tree	Random Forest	CRF
Accelerometer	0.59	0.65***	0.37
Audio	0.55	0.67***	0.31
Accelerometer + Audio	0.64	0.73***	0.42

*** means statistically significant difference at $p < 0.01$

vector. Table 4 provides the global F-measure obtained by adding the location. Apart for CRF, it can be noticed that performances have increased for all feature configurations. Thus the context-aware AR performs better than classical AR without location information. Statistical test between these two experiments for the Accel+Audio feature set revealed a significant difference ($p < 0.01$). Once again, Random Forest has the highest score and it is significantly better than Decision Tree in all conditions while CRF is worse than decision tree. Statistical tests still show a significant impact of audio features on the classification and once again, Random Forest performs better with audio feature set than with accelerometer one.

Table 4. Overall F-measure of activity classification for each Model and feature set with ground truth smartphone location

Feature Set	Decision Tree	Random Forest	CRF
Accelerometer + location	0.64	0.69***	0.38
Audio + location	0.60	0.71***	0.31
Accelerometer +Audio+ location	0.68	0.76***	0.42

*** means statistically significant difference at $p < 0.01$.

5.3 Inferring Smart-phone Context

Since the smartphone location has a positive impact on AR, it was tested whether smartphone location can be inferred from the data. In this experiment, smartphone location was used as class to assess the feasibility of building a location predictor. Results are presented Table 5. Once again, Random Forest performed significantly better than Decision Tree. Unlike previously reported experiments, CRF performance is very close to these of the two other classifiers. It can be seen that very good performance can be reached with Accel+Audio data (F-measure of 91% for Random Forest). Similarly to previous experiments, performances with audio feature set exceed these with accelerometer feature set, for Decision Tree and Random Forest. Moreover, apart for CRF, all classification results for Accel+Audio feature set were significantly superior to results with Accel feature

set only. This means that audio features might have a high predictive power for smartphone location.

Table 5. Overall F-measure of smartphone location classification for Decision Tree, Random Forest and CRF Models and feature set.

Feature Set	Decision Tree	Random Forest	CRF
Accelerometer	0.79	0.84***	0.83
Audio	0.80	0.89***	0.69
Accel+Audio	0.86	0.93***	0.86

*** means statistically significant difference at $p < 0.01$.

5.4 Inferred Context-aware Human Activity Recognition

We complete our experiments by launching again the human activity recognition task with feature sets including smartphone location inferred from a first classifier. Here, we cascaded a smartphone location classifier with a human activity classifier. In order to avoid over-fitting, a portion of the dataset (20 %) was used for training the smartphone location classifier which was then used to predict location on the remaining dataset. The resulting dataset (enhanced with inferred location) was then used for activity recognition classifier in a 10-fold cross-validation way. Table 6 summarizes results (because CRF performed worse than DT for this AR task, it was not used for this experiment). Results are inferior to those in Table 4 (which is not surprising since here location is inferred while it was groud truth on previous experiment) and slightly improved when compared to Table 3. Results should be analysed with caution since both classifiers were trained on reduced datasets for reasons explained before.

Table 6. Overall F-measure of activity classification for Decision Tree and Random Forest and feature set with inferred smartphone location.

Feature Set	Decision Tree	Random Forest
Accelerometer + inferred location	0.60	0.65***
Audio + inferred location	0.56	0.67***
Accel+Audio + inferred location	0.65	0.74***

*** means statistically significant difference at $p < 0.01$.

6 Discussion

One of the main findings of the study is the interest of audio features both for AR and smartphone location recognition. Regarding smartphone location, audio features alone make it possible to overcome the model with accelerometer

features alone (F-measures of 89% for audio vs 84% for accelerometer). But, as shown in Table 5, the most interesting outcome is that audio and accelerometer features are very complementary. Indeed, the dataset with full features shows an increase of Recall for 'hand' (93%) and 'pocket' (91%) well above the one obtained in audio alone (hand: 89%, pocket: 84%) and accelerometer alone (hand: 86%, pocket: 83%). For activity classification, audio features are competitive since results are very similar to the ones with accelerometer feature sets, and even exceed them with Random Forest. However, it must be emphasized that accelerometer features bring far better results with the running and jumping activities than audio features do, which is not surprising. Despite this advantage, the fusion of both feature sets shows again the complementariness of the audio and accelerometer features since performances always significantly increase in that case for AR. However, some subsequent analyses are needed to understand which exact part of the location and activity acoustic signal captures. In particular, a finer feature selection might optimize the current feature set.

Another important finding of the study is the role of the context for improved AR. In this work, the location was simply used as an input feature. If this has brought improvement in all cases, further research must be investigated to know whether location information can be used later in the process. For instance, it might be interesting to train classifiers for every considered location and to use a voting strategy to improve the classification accuracy [17]. In any case, location inference is an important element of the context that can be useful in many applications others than activity recognition.

Three models were tested for AR including a sequence model. Overall, the best model was by far the Random Forest which always showed significant best performance for AR and location classification. RF has showed superiority over decision tree in numerous tasks and is known to handle particularly well imbalanced data sets. Therefore, these results confirms recent studies of the community in similar tasks [8]. The less performing model was Conditional Random Fields that exhibited poor performance for the AR task. It was however competitive for the SP location task. CRF is known to require a high amount of data for training and in the location task only 3 classes with balanced number of instances are provided (about 8400 instances per classes) while in the AR task 9 classes with unbalanced number of instances are provided (from 8400 to 300 instances per class). This suggests that with an increased number of data, CRF could exhibit much better performances.

The data collection also revealed some limits of the smartphone sensor data acquisition. Indeed, we noticed that some smartphones block the acquisition when screen is off. For the data collection, we had to use a screen locker to keep it on and to allow the recording to continue. Thus, real on-line AR system must find solution to react to this unexpected behaviour if they plan to be largely deployed.

The study reported good performance using accelerometer and audio data. These two channels were selected because they are always present on modern

smartphones. However, there are opportunities to improve the classification accuracies even more. For instance, it would be interesting to investigate the data that come from other sensors or phone logs when available. Although data from every possible sensor and phone logs were collected, only accelerometer and microphone were used. For instance, hand location could be inferred from a combination of different clues: a phone call event, no headset or bluetooth device is plugged and motion detected through accelerometer data. Using this kind of information could increase activity recognition accuracy and also diminish the battery usage.

7 Conclusion

The first contribution of this paper is the corpus that was used for the task of activity recognition. Collected from 4 smartphones worn by 19 participants, the data contains both accelerometer and audio samples labeled with 8 different basic physical activities and 3 smartphone locations on volunteers. Moreover, the data was acquired in a sequential way scheduled by scenarios. To the best of our knowledge, this is the first dataset made available with such features.

Beyond the corpus, the main contribution is the importance of audio features for the AR task. The audio feature set proved to be competitive with the accelerometer feature set for the AR task and even outperformed it with some of the 3 state of the art classifiers used. Moreover, the best performance was achieved when combining both feature sets together.

Finally, our study highlights the importance of smartphone location. Our approach includes it as an attribute of the feature set. This resulted in an increase of AR performances which confirmed our assumption. Yet, smartphone location information can be integrated in other ways in such a system of recognition (e.g., classifiers are trained for different locations and a voting strategy selects one).

These results will be useful for the ambient intelligence community by shedding light on the role of audio features for AR as well as on the importance of modelling the smartphone context for the applications. Nevertheless, we plan to undertake more studies to assess the interest of sequence models and to collect data "in the wild" to validate the approach in real ecological situations.

References

1. BE Ainsworth, WL Haskell, SD Herrmann, N Meckes, DR Bassett, C Tudor-Locke, JL Greer, J Vezina, MC Whitt-Glover, and AS Leon. 2011 compendium of physical activities: a second update of codes and met values. *Medicine and science in sports and exercise*, 43(8):1575–1581, 2011.
2. Khaled Alanezi and Shivakant Mishra. Impact of smartphone position on sensor values and context discovery. `http://digitool.library.colostate.edu/exlibris/dtl/d3\_1/apache\_media/L2V4bGlicmlzL2R0bC9kM18xL2FwYWNoZV9tZWRpYS8yMTIyNjM=.pdf`.
3. Ling Bao and Stephen S. Intille. Activity recognition from user-annotated acceleration data. In *Pervasive*, pages 1–17, 2004.

4. David Blachon, François Portet, Laurent Besacier, and Stéphan Tassart. RecordMe: A Smartphone Application for Experimental Collections of Large Amount of Data Respecting Volunteer's Privacy. In R. Hervás et al., editor, *UCAmI 2014*, pages 345–348, 2014.
5. Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.
6. Yohan Chon, Elmurod Talipov, and Hojung Cha. Autonomous management of everyday places for a personalized location provider. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, 42(4):518–531, 2012.
7. Joëlle Coutaz, James L. Crowley, Simon Dobson, and David Garlan. Context is key. *Communications of the ACM*, 48(3):49–53, 2005.
8. Božidara Cvetković, Boštjan Kaluža, Radoje Milić, and Mitja Luštrek. Towards human energy expenditure estimation using smart phone inertial sensors. In *Ambient Intelligence*, pages 94–108, 2013.
9. Irina Diaconita, Andreas Reinhardt, Frank Englert, Delphine Christin, and Ralf Steinmetz. Do you hear what i hear? using acoustic probing to detect smartphone locations. In *Pervasive Computing and Communications Workshops (PERCOM Workshops), 2014 IEEE International Conference on*, pages 1–9. IEEE, 2014.
10. Nathan Eagle and Alex Pentland. Reality mining: sensing complex social systems. *Personal and ubiquitous computing*, 10(4):255–268, 2006.
11. Ozlem Durmaz Incel, Mustafa Kose, and Cem Ersoy. A review and taxonomy of activity recognition on mobile phones. *BioNanoScience*, 3(2):145–171, 2013.
12. Mustafa Kose, Ozlem Durmaz Incel, and Cem Ersoy. Online human activity recognition on smart phones. In *Workshop on Mobile Sensing: From Smartphones and Wearables to Big Data*, pages 11–15, 2012.
13. John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, pages 282–289, 2001.
14. Emiliano Miluzzo, Michela Papandrea, Nicholas D Lane, Hong Lu, and Andrew T Campbell. Pocket, bag, hand, etc.-automatically detecting phone context through discovery. *Proc. PhoneSense 2010*, pages 21–25, 2010.
15. Jun-geun Park, Ami Patel, Dorothy Curtis, Seth Teller, and Jonathan Ledlie. Online pose classification and walking speed estimation using handheld devices. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, pages 113–122. ACM, 2012.
16. J. Ross Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1993.
17. Sasank Reddy, Min Mun, Jeff Burke, Deborah Estrin, Mark Hansen, and Mani Srivastava. Using mobile phones to determine transportation modes. *ACM Transactions on Sensor Networks (TOSN)*, 6(2):13, 2010.
18. M.R Segal. Machine learning benchmarks and random forest regression. Technical report, University of California, 2004.
19. Charles Sutton and Andrew Mccallum. *Introduction to Conditional Random Fields for Relational Learning*. MIT Press, 2006.
20. T.L.M. van Kasteren, G. Englebienne, and B.J.A. Kröse. An activity monitoring system for elderly care using generative and discriminative models. *Personal and Ubiquitous Computing*, 14(6):489–498, 2010.
21. Jun Yang. Toward physical activity diary: motion recognition using simple acceleration features with mobile phones. In *Proceedings of the 1st international workshop on Interactive multimedia for consumer electronics*, pages 1–10. ACM, 2009.